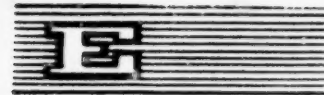




UNITED NATIONS
ECONOMIC AND SOCIAL COUNCIL



Distr.
LIMITED

ST/ECA/SM/11
9 September 1981

Original : ENGLISH

ECONOMIC COMMISSION FOR AFRICA
Seminar on Household Surveys
Addis Ababa, 21-30 September 1981

LC/NAIROBI

ON SURVEY DATA ANALYSIS

CONTENTS

	<u>Paragraphs</u>	<u>Pages</u>
I. INTRODUCTION	1 - 6	1
II. SOME POLICY QUESTIONS	7 - 17	1 - 2
III. THE EXPLORATION OF DATA STRUCTURES AND MODEL FITTING	18 - 27	3 - 8
General	18 - 22	3
Analysis of Contingency Tables	23	3 - 4
Cluster Analysis	24	4 - 5
Aid Analysis	25	5 - 6
Box-Jenkins Analysis	26	6
Model Fitting	27	6 - 8
IV. BIBLIOGRAPHY	-	9 - 11

I. INTRODUCTION

1. A paper on survey data evaluation and analysis (E/CN.14/SM/27) was produced for the second meeting of the working group on organization, content and methodology of household surveys in October 1979. In the last meeting of the working group (June-July 1981) another paper on survey data analysis was presented (E/CN.14/SM/39).
2. The last paper dealt with analytical requirements and responsibilities, activities prior to analysis, analytical options, the data base approach and some notes on analysis. Rather little was said about the analytical tools for analyzing survey data.
3. The main purpose of the present paper is to give a brief account of a series of techniques for the exploration of data structure and analysis of survey data. A bibliography of relevant books is included. Completeness has not been attempted. One reason is that there is no particular limit to the scope and applicability of statistical theory and method in this connection.
4. The working group has earlier strongly stressed the importance of a close cooperation between statisticians and users of statistics in order to be able to formulate the statistical problems correctly and to present and analyse the data so as to be useful for decision making and action taking. This topic is no doubt of utmost importance and should be dealt with in some detail when discussing the AHSCP.
5. While recognizing the importance of proper problem formulation, it is believed that there is a strong connection between the success achieved in this area and knowledge of statistical tools available for describing and analyzing survey data. There is a two-way process.
6. A prerequisite for applying techniques of exploring data structures and analysis of data is that the previous stages of survey work has been well handled. It means that added importance is attached to an adequate sampling design, control of non-sampling errors, etc.

II. SOME POLICY QUESTIONS

7. The question if producers of statistics should carry out consumer-oriented analysis of official statistics has been much discussed in recent years. This theme was for example a topic of discussion at the ISI-session in Warszawa in 1975. The question has also been discussed at several meetings of the UN economic commissions.
8. It has been generally agreed that the producers of statistics should carry out producer-oriented analysis. This includes analysis of data quality, calculations of sampling errors, etc. The purpose with this type of analysis is to reveal limitations concerning the data collected and to warn the users from drawing wrong conclusions. When it comes to consumer-oriented analysis - which is the subject of this paper - different opinions have been expressed. However, the current trend is that the users of statistics request more elaborate statistical products as a basis for decision-making.
9. It is believed that a statistical office is the proper place for the presentation and analysis of survey data, provided that there is a close cooperation between statisticians and planners at various stages of the survey work. One obvious reason is the statistical nature of the techniques applied as well as

possession of first hand knowledge of the quality of the data collected. Of course, this should not prevent users of data to get engaged in analytical work as well.

10. The duty of the statistical office to carry out consumer-oriented analysis has a series of consequences both with regard to statistical training and to statistical organization. There are also budgetary consequences, as more resources have to be allocated to the planning and analysis of surveys.

11. The importance of training can hardly be exaggerated. Multivariate analysis is usually not extensively dealt with at lower academic levels. Accordingly, post-graduate courses have to be organized to meet the needs of education in this discipline of statistics.

12. In addition to university training it appears necessary to organize in-service training in the central statistical offices. This in turn calls e.g. for strengthening of the statistical libraries, which should have an adequate supply of textbooks and journals in the field of statistical methodology.

13. For each survey there should be a project officer, who is involved at all stages, from the start and up to the writing of reports. His first duty should be to work closely with planners and users of the survey results to make the reports consumer-oriented. He will have to be backed up at various stages by specialists in fields such as administration, organization, sampling, data processing, statistical analysis, etc. However, he should himself have a broad general knowledge of various areas, particularly methods of statistical analysis.

14. It is advantageous to set up a section, consisting of mathematically oriented statisticians who are specialists in fields such as sampling and analysis. Their task should be to assist the project officers, to study continuously new statistical methodology and to arrange in-service training for other statisticians in the office.

15. Concerning the budgetary aspects it should be kept in mind that the additional cost of more advanced description and analysis of household surveys is small in comparison with the other costs of interviewing, transport, checking, coding etc. The marginal utility is high. No doubt statistics will be more highly appreciated as a consequence of analytical work.

16. Access to computers is fundamental in the analysis of surveys. There are a great many computer programs for the purpose. Attempts should be made to build up useful program libraries in the statistical offices. It is suggested that the ECA makes a contribution in this respect.

17. It would be valuable if an African Statistical Journal could be set up with support from the ECA, containing e.g. articles on applications of statistical analysis of survey data in the African context. This would certainly promote the African household survey programme.

III. THE EXPLORATION OF DATA STRUCTURES AND MODEL FITTING

18. Until recently, the almost only type of sample survey analysis consisted of estimations of population means, totals, frequency distributions, maybe sampling errors, correlation coefficients, cross-tabulations, graphs and the like. If well done and carefully examined, much information can be obtained from such exercises, also with regard to relationships between variables.

19. Particularly two factors have transformed the analysis of survey data in the last decade: the development of new techniques for data analysis and the availability of powerful computers to survey analysts.

20. Until rather recently statistical methods for dealing with relationships between variables have been developed to meet the needs of the experimental scientist. The social and economic research workers on the other hand, have mostly to use surveys, which result in a complex and often large set of data. There is a need for exploratory techniques, which adequately describe structural features in the data set. This may form a prelude to theory generation and model construction.

21. Many writers on the subject distinguish between techniques which aim at a more refined form of data description (exploratory techniques) on the one hand and techniques which aim at showing causal relationships (model fitting) on the other hand. The distinction may sometimes be a bit difficult to make. Anyway, in view of the relative scarcity of survey data in the African region the techniques indicated in the following mainly belong to the exploratory techniques. Only one paragraph deals with model fitting explicitly.

22. The choice of statistical techniques, which are briefly presented here, is selective. The same goes for the bibliography given. The aim of the paper is rather to draw attention to the importance of considering modern methods of analysis when discussing the African household survey programme. Practical applications will certainly give interesting experiences and add more substance to the programme.

23. Analysis of contingency tables. This type of analysis is mainly concerned with methods of dealing with frequency data occurring in the form of cross-classifications or contingency tables, such as 2x2 contingency tables, rxc contingency tables or multidimensional tables. When a population is classified into several categories one can count the individuals in each category. These counts or frequencies are the type of data with which this kind of analysis is primarily concerned. The analysis is dealing with qualitative rather than quantitative data obtained from measurement of continuous variables. Of course, continuous data can often be put into discrete form by the use of intervals on a scale.

It is to a large extent a matter of hypothesis testing techniques for the analysis of contingency tables. Among these may be mentioned the chi-square test, Fisher's exact test for 2x2 tables, Mc Nemar's test for correlated proportions in a 2x2 table, Gart's test for order effect, the $V(x^2)$ method when combining information from several 2x2 tables, Cochran's method for a similar situation, Lancaster and Irwin's method for partitioning rxc tables, etc.

Frequently one may be interested in measuring the strength of association between the qualitative variables involved in addition to testing the significance. A number of measures have been constructed, such as the mean square contingency coefficient $\phi^2 = x^2/N$, Pearson's coefficient of contingency

$$P = \sqrt{\left(\frac{x^2/N}{1+x^2/N}\right)}, \text{ Kendall's and Stuart's } T = \frac{x^2/N}{\sqrt{(r-1)(c-1)}}$$

$$\text{Cramer's } C = \frac{x^2/N}{\min(r-1, c-1)}, \text{ Goodman and Kruskal's } \lambda \text{ and Gamma measures and Somer's } d\text{-measure.}$$

Another approach has also been considered, namely that of fitting models and of estimating the parameters in the models. The starting-point is a theory about the observations and the parameters in the model represent the effects that particular variables or combinations of variables have in determining the values. Of particular interest are methods developed for fitting log-linear models to multidimensional frequency data and the description of these data in parametric terms. Among statisticians who have been engaged in developing the methods in question may be mentioned Bartlett, Bock, Haberman, Newton, Raphson, Deming, Stepham, Birch, Bishop, Fienberg and Goodman. The most comprehensive account is given in Bishop, Fienberg and Holland (1975) : Discrete Multivariate Analysis. Another book dealing with the subject is Everitt (1977) : The Analysis of Contingency Tables. Both volumes are included in the bibliography given. Finally it may be mentioned that a number of computer programs for the fitting of log-linear models to complex multidimensional contingency tables are available.

24. Cluster analysis. This is a label for a number of different numerical methods of classification. A common feature is that these methods - based on some criterion of similarity - group variables or objects in clusters. The units in a certain cluster should be as homogenous as possible while the clusters between each other should be as different as possible.

Several measures of similarity are described in e.g. Sneath and Sokal : Principles of Numerical Taxonomy. The most common measures are the Euclidean distance and the correlation coefficient. Measures have been defined for quantitative as well as qualitative variables.

Hierarchical techniques are probably the most commonly used class of clustering methods. They may be subdivided into agglomerative methods which proceed by a series of successive fusions of individuals into groups and divisive methods which partition a set of n individuals successively into finer partitions. The results of both agglomerative and divisive techniques may be presented in the form of a s.c. dendrogram, which is a two-dimensional diagram illustrating the fusions or partitions which have been made at each successive level.

Another class of clustering techniques are those which seek to partition the set of data so as to optimize some predefined clustering criterion. These techniques assume that the number of groups into which to partition the data has been set by the user to some value, say k. Many clustering criteria have been suggested. Probably the most common is minimization of the pooled within cluster sum of squares, a method proposed by e.g. Singleton and Kantz, Mac Queen and Ball and Hall.

Cluster analysis can be utilized for descriptive purposes by meaningful grouping of a large number of individual data. It can also be used to explore the data structure to obtain a basis for model building or for hypothesis generating purposes. When evaluating the results of cluster analysis it is important to make use of available knowledge of the subject-matter.

A common problem when applying cluster analysis is that the variables used may have been measured by different types of scales. Then one has to decide upon some specific type of scale and then transform all the variables into that scale. Of course, this leads to a certain loss of information.

The number of papers describing applications of clustering techniques is considerable. There have been several reported applications in psychiatry, psychology, education and market research. Cluster analysis has been used for establishing a nomenclature for occupations and for studying change of occupations. It has also been used to study areas within cities. Finally it may be mentioned that cluster analysis has been used to form a basis for sampling design, e.g. as a stratification tool using as a criterion of effectiveness the ability to increase the efficiency of estimation.

There are many packages of computer programs available for various methods of cluster analysis.

In addition to the volume by Sneath and Sokal may be mentioned Ball: Classification Analysis, Everitt: Cluster Analysis, Jardine and Sibson: Mathematical Taxonomy, Singleton and Kantz: Minimum Square Error Clustering Algorithm and various publications by Wolfe, e.g. NORMIX: Computational Methods for estimating the parameters of Multivariate Normal Mixtures of Distributions. Reference is made to the bibliography attached.

25. AID Analysis. AID means Automatic Interaction Detection, which indicates that the method is designed to detect interactions in a data set. AID was introduced in 1964 by Morgan and Sonqvist at the University of Michigan. They also made ADP-programs. As an analytic method AID has developed along with the program, which is characteristic of other techniques of exploring data structures, such as cluster analysis. It is usually impossible to manage the calculations without the use of dator capacity.

A group of n individuals is under study, each characterized by a set of variables $(y, y_2 \dots y_m; x_1, x_2, \dots x_p)$. The y_i are taken to be the dependent variables, x_j independent or predictor variables. In some contexts there is just one dependent variable y . AID is a technique of step by step binary segmentation. Firstly, the sample as a whole is split into two subgroups and the same procedure is applied sequentially to resulting subgroups. However, at any stage decisions must be made on the particular predictor variable to be used in inducing the split, and how the two subsets of its values are to be formed. Since the objective is to explain the variation in the dependent variables, it seems sensible to make these two decisions so that the resulting subgroups are as different as possible and are internally as homogeneous as possible with regard to the y_i 's.

The basic strategy is to examine each predictor variable x_j and to choose a partition of its set of values such that the induced binary split results in subgroups that maximize the chosen "distance" function. The next step is to choose among these the variable which produces the overall maximum. In a sense it is the best explanation of the y_i one can get through a single split into two groups. The process continues by sequentially splitting the subgroups obtained from the previous stages using the same criterion for deciding which predictor to use. Rules have been given for deciding when to stop the process.

The result of the binary segmentation process is a hierarchical tree. This is a diagram, which gives information on the terminal or final groups, association between predictors and interaction effects.

The technique has been widely applied, e.g. for studies of salaries, prices, rents, effects of television broadcast, smoking, problem drivers, emigration, education etc. It has also been used as a tool for planning sample surveys.

Particular reference is made to Sonquist : Multivariate Model Building. The Validation of a Search Strategy, Sonquist, Baker and Morgan : Searching for Structure and Morgan Messenger : THAID. A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables.

26. Box-Jenkins Analysis. This technique refers to a special strategy analysis. Where the main purpose is to give short forecasts of time series. The method can more generally be used for description and analysis of time series. One might for instance test some hypothesis concerning changes of trends as a consequence of external events, such as strikes, new regulations or energy crises.

The Box-Jenkins method calls for extensive numerical calculations and it is therefore necessary to have access to dator capacity and special ADP-programs to carry out the analysis.

Box-Jenkins analysis can be carried out at two levels. The first is the level of univariate models, when the analysis is based only on the historical development of specific series. The results are used to eliminate seasonal effects and to identify points, where the quality of data may be defective. The second level is that of building up transfer function models taking into account dependent series as well as a number of explanatory series. Transfer function models may be used for a) short forecasts b) improvement of preliminary statistics with the help of leading indicators c) estimation of the effects of internal/external events and perhaps making corrections for these effect.

The method has been utilized in many contexts, e.g. improvement of data quality, improvement of preliminary indices, studies of the effect of the oil crises on the consumption of electricity by the households, etc.

Reference is particularly made to Box and Jenkins : "Time Series Analysis" and Jenkins: "Practical Experiences with Modelling and Forecasting Time Series".

27. Model Fitting. This term relates to the situation when one has reached an insight into the causal relationships between variables in some subject area. Thus a model of the form $(y_1, y_2, \dots, y_m) = f(x_1, x_2, \dots, x_p) + E$ can be set up where the functional relationships are known. This make it possible not only to fit a model, but also to test the goodness of fit of the model to the observed data.

Of course a model is only an approximation to the real world. It is usually impossible to include all variables relevant to the problem dealt with, which is reflected in the size of the error term ϵ .

The type of models most commonly used in survey analysis is the class of linear models. A dependent variable is assumed to be a linear function of the parameters of the model. The relationship is $f(y) = \sum_{j=1}^p \beta_j x_j + \epsilon$ where $f(y)$ is a dependent variable, β_j 's are the parameters and the x_j 's are the explanatory variables. Interactions can be built into the model by defining new variables which represent these interactions. This is a problem which has led to the development of search techniques, some of which have been indicated earlier in this paper.

The most commonly used form of the general linear model is ordinary least squares regression. Additional assumptions must be made about the distribution of the ϵ 's if the parameters are to be estimated optimally. The disturbances must be uncorrelated with each other and with the x 's and must all have equal variance. The analysis of variance is a special case of multiple regression which arises when all the explanatory variables are nominal-scale or categorical variables, while the dependent variable is still a continuous interval-scale variable. The analysis of covariance may be used when some of the explanatory variables are continuous and some are nominal.

Frequently one has to assume several variables which have to be considered simultaneously. Two ways of dealing with such situations are multiple regression and simultaneous equation estimation. The difference between these two methods is that in the latter case there will be more than one dependent variable. For simultaneous equation estimation a variable may appear as dependent in one equation and as an explanatory variable in others. This technique has been extensively used in the field of econometrics.

Canonical correlation is another technique that represents an extension of linear-regression analysis. It is concerned with two sets of variables. The method comprises the derivation of a variate - a linear combination of the variables in a set - for each set such that the correlation between the sets are maximized. To date, applications of the method have been principally in economics for the estimation of demand and supply relationships, where both price and quantity may be expressed as a linear combination of variables.

The basic hypothesis of discriminant analysis is that a population is assumed to be made up of two distinct subpopulations. It is further assumed that it is possible to find a linear function of certain measures or attributes of the population that will allow an observer to discriminate between the two subpopulations. Discriminant analysis - particularly developed by Fisher, Neyman and Pearson - was originally used in the biological sciences. However, the technique has many possible applications in the social sciences. It has for instance been occasionally used in the field of transport statistics.

Factor analysis comprises a set of widely differing techniques which have been developed since Spearman introduced the conceptual framework in the beginning of the 20th century. Principal component analysis, on the other hand, is a more well-defined procedure. Basically these methods attempt to condense a set of variables, that is, the factors or components, which have not been directly measured. The methods have their roots in the basic principles of multivariate linear-regression analysis and there has been some controversies as to the nature of the methods. Kendall has suggested that principal component analysis is the search through data to try to find components that may reduce the dimensions of variation while

factor analysis starts with the hypothesis of a model and tests it against the data to see if it fits. Factor analysis was originally developed for use in the field of psychology but has later on been applied in various other fields. It is not sufficient to obtain just a mathematical solution. The solution should also be conceptually interpretable. Frequently this is one of the major difficulties encountered in factor analysis.

Two other methods of model-fitting may be briefly mentioned. Path analysis has become one of the standard tools for analysing interval-level data for which an explicit causal structure can be stipulated. Probit-and logit analysis represent techniques to be used in a situation when one wishes to estimate the probability of some occurrence as a function of a set of independent variables that may take values over a wide range.

There is a vast literature on the various model-fitting techniques.

IV. BIBLIOGRAPHY

- Anderberg, M.R. (1973) ; Cluster Analysis for Applications (Head Press, New York)
- Andersson, T.W. (1958) : An Introduction to Multivariate Statistical Analysis (Wiley, New York)
- Andrews, F.M and Messenger, R.C. (1973) : Multivariate Nominal Scale Analysis (Inst. for Social Research, Univ. of Michigan)
- Armitage, P. (1971) : Statistical Methods in Medical Research (Blackwell Scient. Publ., Oxford, Edinburgh)
- Ball, G.H. (1971) : Classification Analysis (Stanford Research Institute SRI Project 5533)
- Beale, E.M.L. (1969): Cluster Analysis (Scientific Control Systems, London)
- Bibby, J. and Toutenburg, H.(1977) : Prediction and Improved Estimation in Linear Models (Wiley, London)
- Bishop, Y.M., Fienberg, S.E. and Holland, P.W. (1975): Discrete Multivariate Analysis; Theory and Practice (The M.I.T. Press, Cambridge, Massachusetts)
- Borgatta, E.F. (1969) : Sociological Methodology 1969 (San Francisco : Jossey-Bass)
- Box, G.E.P. and Jenkins, G.M. (1976) : Time Series Analysis. Forecasting and control. (Holden Day)
- Cassel, C-M., Sarndal, C-E. and Wretman, J.H. (1977) : Foundations of Inference in Survey Sampling (Wiley, New York)
- Cattell, R.B. (1952) : Factor Analysis (Harper, New York)
- Cox, D.R. (1970): Analysis of Binary Data, (Methuen, London)
- Cramér, H. (1946) : Mathematical Methods for Statistics (Princeton Univ. Press)
- Davis, D. : A factorial ecology of Malmö 1965; the social geography of a city (Lunds Universitets Geografiska Institution, Lund, Sweden)
- Davis, J.A. (1971) : Elementary Survey Analysis (Englewood Cliffs, N.J. : Prentice Hall)
- Dhrymes, P.J. (1970) : Econometrics : Statistical Foundations and Applications (Harper and Row, New York)
- Draper, N.R. and Smith, H. (1966) : Applied Regression Analysis (Wiley, New York)
- Everitt, B.S. (1974) : Cluster Analysis (Heinemann Press, London)
- Everitt, B.S. (1977) ; The Analysis of Contingency Tables (Chapman and Hall, London)
- Finney, D.J. (1971) : Probit Analysis (Univ. Press, Cambridge)
- Fisher, R.A. (1950) : Statistical Methods for Research Workers (Oliver and Boyd, Edinburgh)
- Fleiss, J.L. (1973) : Statistical Methods for Rates and Proportions (Wiley, New York)
- Frankel, M.R. (1971) : Inference from survey samples (Inst. for social research, Univ. of Michigan)
- Godambe, V.P. and Sprott, D.A. (1971) : Foundations of Statistical Inference (Symposium on the Foundations of Statistical Inference held at the Department of Statistics, Univ. of Waterloo, Canada)

- Goldberger, A.S. (1968) : Topics in Regression Analysis (Collier-Macmillan, London)
- Goldberger, A.S. and Duncan, O.D. (1973) : Structural Equations Models in the Social Sciences (Seminar Press, New York)
- Gorsuch, R.L. (1974) : Factor Analysis (W.B. Saunders Co, Philadelphia)
- Haberman, S.J. (1974) : The Analysis of Frequency Data (Univ. of Chicago Press, Chicago)
- Hags, W.L. (1973) : Statistics for Social Sciences (Holt, Reinhort and Winston, New York)
- Hanushek, E.A. and Jackson, J.E. (1977) : Statistical Methods for Social Scientists (Academic Press, New York)
- Harman, H.H. (1960) : Modern Factor Analysis (Univ. of Chicago Press, Chicago)
- Hollander, M. and Wolfe, D.A. (1973) : Non-Parametric Statistical Methods (Wiley, New York)
- Hope, K. (1968) : Methods of Multivariate Analysis (Univ. of London Press, London)
- Jardine, N. and Sibson, R. (1971) : Mathematical Taxonomy (Wiley, London)
- Jenkins, G.M. (1979) : Practical experiences with Modelling and Forecasting Time Series Jenkins and Partners Ltd)
- Johnston, J. (1972) : Econometric Methods (Mc Graw-Hill, New York)
- Kendall, M.G. (1965) : A course in Multivariate Analysis (Griffin, London)
- Kendall, M.G. and Stuart, A. (1961) : The Advanced Theory of Statistics (Griffin, London)
- Kendall, M.G. (1955) : Rank Correlation Methods (Hafner, New York)
- Kirk, R.E. (1968) : Experimental Design Procedures for the Behavioural Sciences (Brooks & Cole, Calif)
- Kish, L. (1965) : Survey Sampling (Wiley, New York)
- Lawley, D.N. and Maxwell, A.E. (1971) : Factor Analysis as a Statistical Method (Butterworths, London)
- Leser, C.E.V. (1974) : Econometric Techniques and Problems (Griffin, London)
- Lindsey, J.K. (1973) : Inference from Sociological Survey Data (Elsevier, Amsterdam)
- Mac Naughton-Smith, P. (1965) : Some Statistical and Other Numerical Techniques for Classifying Individuals (Home Office Research Unit Report No. 6, London)
- Miller, R.G. Jr. (1966) : Simultaneous Statistical Inference (Mc Graw-Hill, New York)
- Morgan, J.N. and Messenger, R.C. (1973) : THAID, A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables (Inst. for Social Research, Univ. of Michigan)
- Moser, C.A. and Kalton, G. (1971) : Survey Methods in Social Investigations (Heinemann, London)
- Mosteller, F. and Tukey, J.W. (1977) : Data Analysis and Regression (Addison-Wesley, London etc.)
- Namboodiri, N.K. Carter, L.F. and Blalock, H.M. (1975) : Applied Multivariate Analysis and Experimental Designs (McGraw-Hill, New York)
- Neter, J.A. and Wasserman, W. (1974) : Applied Linear Statistical Models (Irvin, Homewood, Illinois)

- O'Muircheartaigh, C. and Payne, C. (1977) : Exploring Data Structures (Wiley, London)
- O'Muircheartaigh, C.A. and Payne, C. (1977) : Model Fitting (Wiley, London)
- Pielou, E.C. (1969) : An Introduction to Mathematical Ecology (Wiley, New York)
- Plackett, R.L. (1974) : The Analysis of Categorical Data (Griffin, London)
- Rao, C.R. (1952) : Advanced Statistical Methods in Biometric Research (Wiley, New York)
- Rao, C.R. (1968) : Linear Statistical Inference (Wiley, London)
- Rao, P. and Miller, R.L. (1971) : Applied Econometrics (Wadsworth Publ. Co, Belmont)
- Rummel, R.J. (1973) : Applied Factor Analysis (Evanston, Ill; Northwestern Univ. Press)
- Saker, G.A.F. (1977) : Linear Regression Analysis (Wiley, New York)
- Searle, S.R. (1971) : Linear Models (Wiley, New York)
- Schuessler, K. (1971) : Analysing Social Data: A Statistical Orientation (Houghton Mifflin, Boston)
- Singleton, R.C. and Kantor, W. (1965) : Minimum Square Error Clustering Algorithm (Res. Inst., Stanford)
- Sneath, P.H.A. and Sokal, R.R. (1973) : Principles of Numerical Taxonomy (Freeman, London)
- Sonquist, J.A. (1970) : Multivariate Model Building (Inst. for Social Research, Univ. of Michigan)
- Sonquist, J.A. Baker, E.L. and Morgan, J.N. (1971) : Searching for Structure (Inst. for Social Research, Univ. of Michigan)
- Sonquist, J.A. and Morgan, J.N. (1974) : The Detection of Interaction Effects (Ann Arbor, Inst. for Social Research)
- Sprent, P. (1969) : Models in Regression and Related Topics (Methuen, London)
- Stopher, P.R. and Meyburg, A.H. (1979) : Survey Sampling and Multivariate Analysis for Social Scientists and Engineers (Lexington Books, Massachusetts)
- Theil, H. (1973) : Principles of Econometrics (Wiley, London)
- Thurstone, L.L. (1947) : Multiple Factor Analysis (Univ. of Chicago Press, Chicago)
- Tintner, G. (1952) : Econometrics (Wiley, New York)
- Tukey, J.W. (1977) : Exploratory Data Analysis (Addison-Wesley, London etc.)
- Van de Geer, J.P. (1971) : Introduction to Multivariate Analysis for the Social Sciences (Freeman, San Francisco)
- Williams, E.J. (1959) Regression Analysis (Wiley, New York)
- Wolfe, J.H. (1967) : NORMIX : Computational Methods for estimating the parameters of Multivariate Normal Mixtures of Distributions (Res. Memo SRM 68-2, San Diego Navy Personnel Research Activity)
- Wolfe, J.H. (1969) : Pattern Clustering by Multivariate Mixture Analysis (Res. Memo SRM 69-19, San Diego: US Navy Personnel Research Activity)
- Tyron, R.C. and Bailey, D.E. (1970) : Cluster Analysis (Mac Gray-Hill, New York)
- Yates, F.S. (1981: Fourth Edition) Sampling Methods for Censuses and Surveys (Griffin, London)